2018 US House Elections Forecast

Nick Ahamed

12 February 2018

# Pooling the polls to forecast the 2018 US House elections

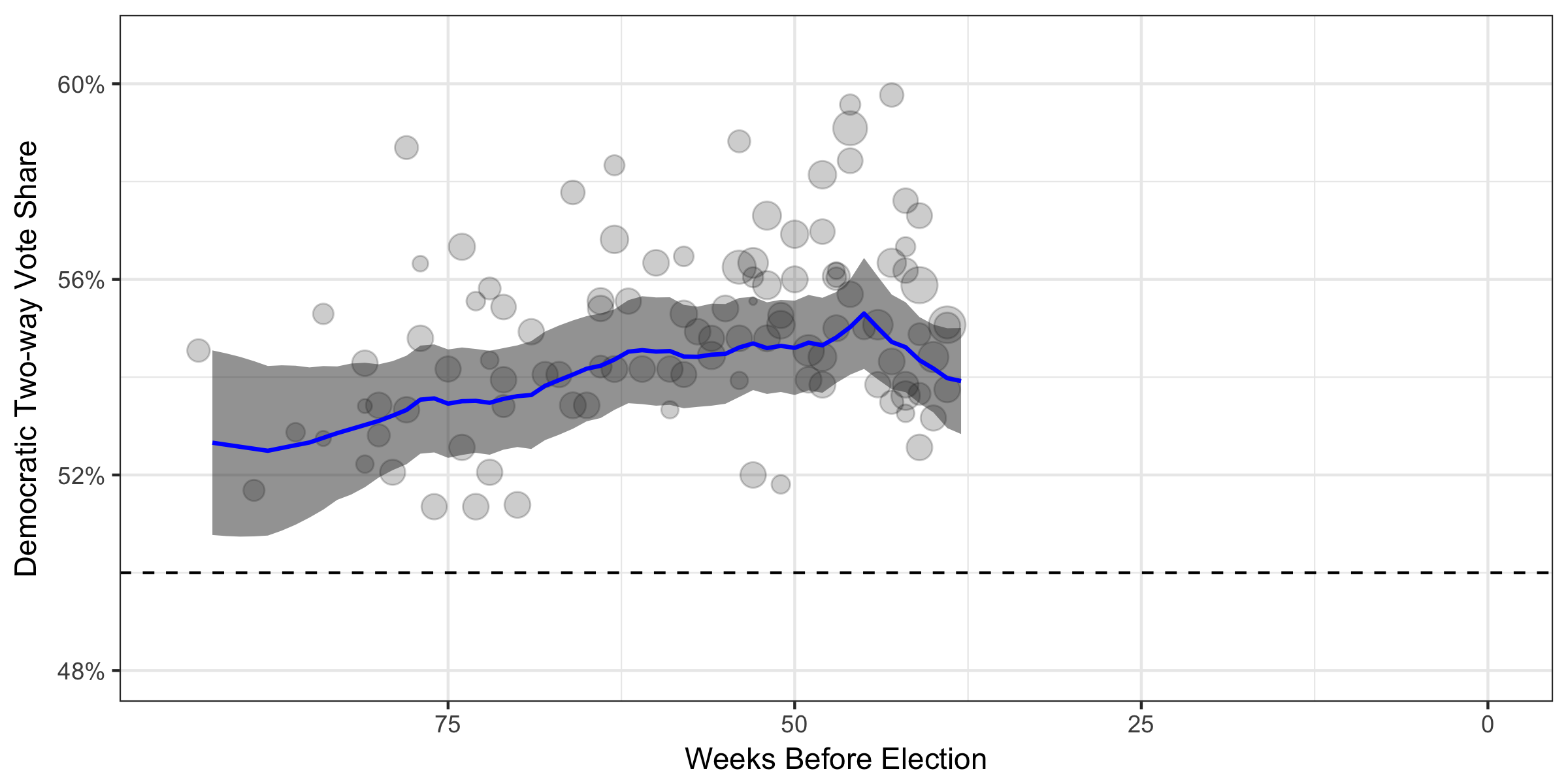
## Introduction

Using previously estimated pollster and sampling universe biases and current polling data, I forecast the results of the 2018 U.S. House of Representatives elections. My Bayesian model predicts Democrats will receive 54% of the popular vote, yielding 225 seats in the House of Representatives, 7 more than the 218 needed to win a majority. Based on current polling, Democrats win a majority of seats on Election Day in 55% of simulations. Lastly, I have identified about 30 battleground districts that would be strategic targets if swing from 2016 is roughly uniform.

## Week-by-week estimate of support

I use a Bayesian random-walk model to estimate the true level of support over time. Based on the most current data, my forecast gives Democrats 53.9% support on the two-way generic ballot, which translates to 225 seats in the House of Representatives. There is relatively little volatility in week-to-week movements. Whereas in the past six election cycles, 95% of week-to-week movement was within 1.5pp, this cycle, 95% of movement is within just 0.7pp. This perhaps reflects strong stability in voter opinion: regardless of the news cycle, most voters are consistent in their support. It is an important reminder that Democrats do not need to respond to the day-to-day and hour-to-hour machinations of Trump, but should focus on longer-term base building.

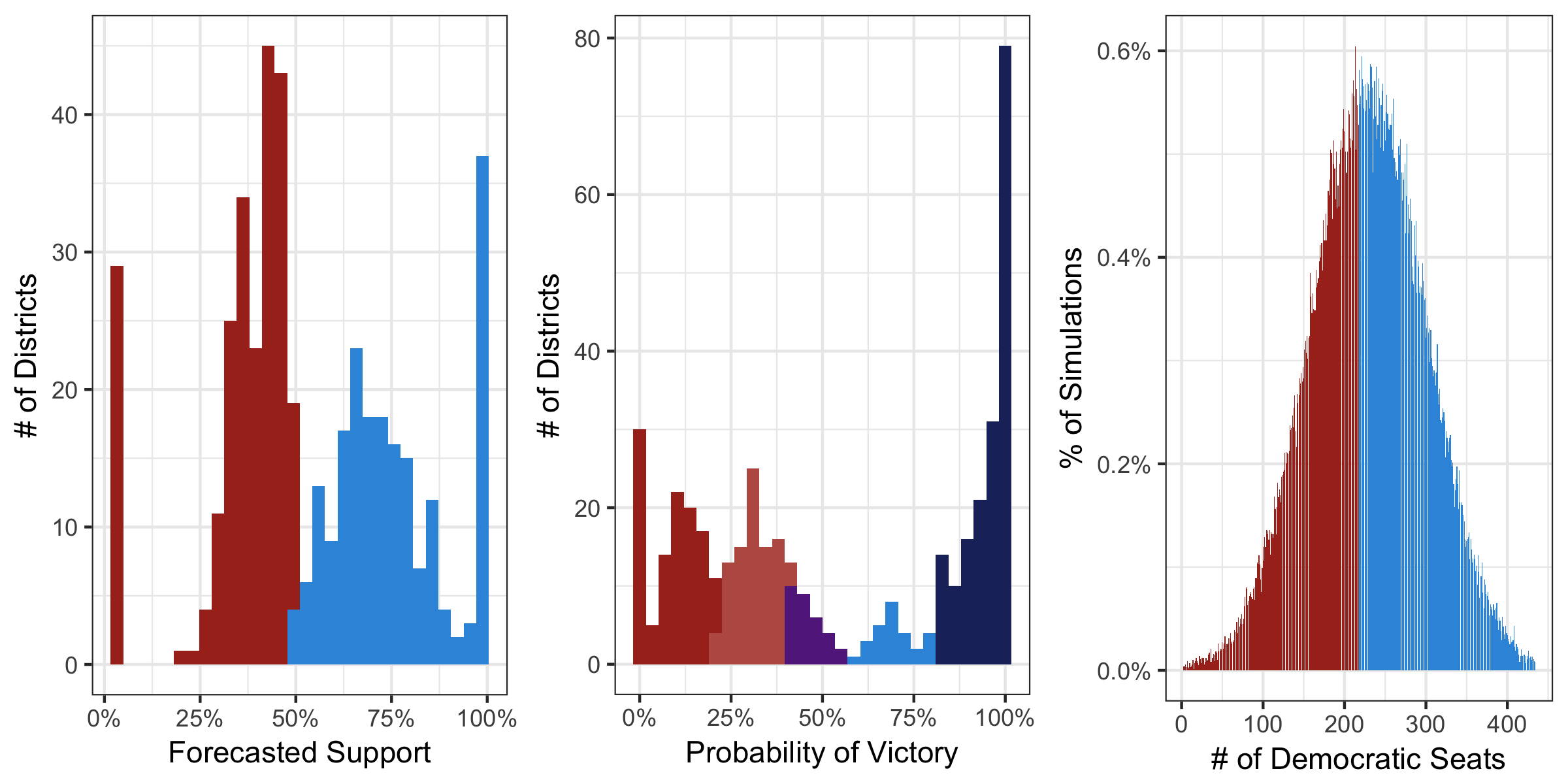
We saw the highest support for Democrats at the end of 2017, possibly associated with the Alabama special election or the tax bill, but there has been a slight slide in support since. Despite this, a +8 two-way margin is a good sign for Democrats. In comparison, at this point in 2010, Republicans polled at 54.2%, predicting they would win 227 seats; they won 242.



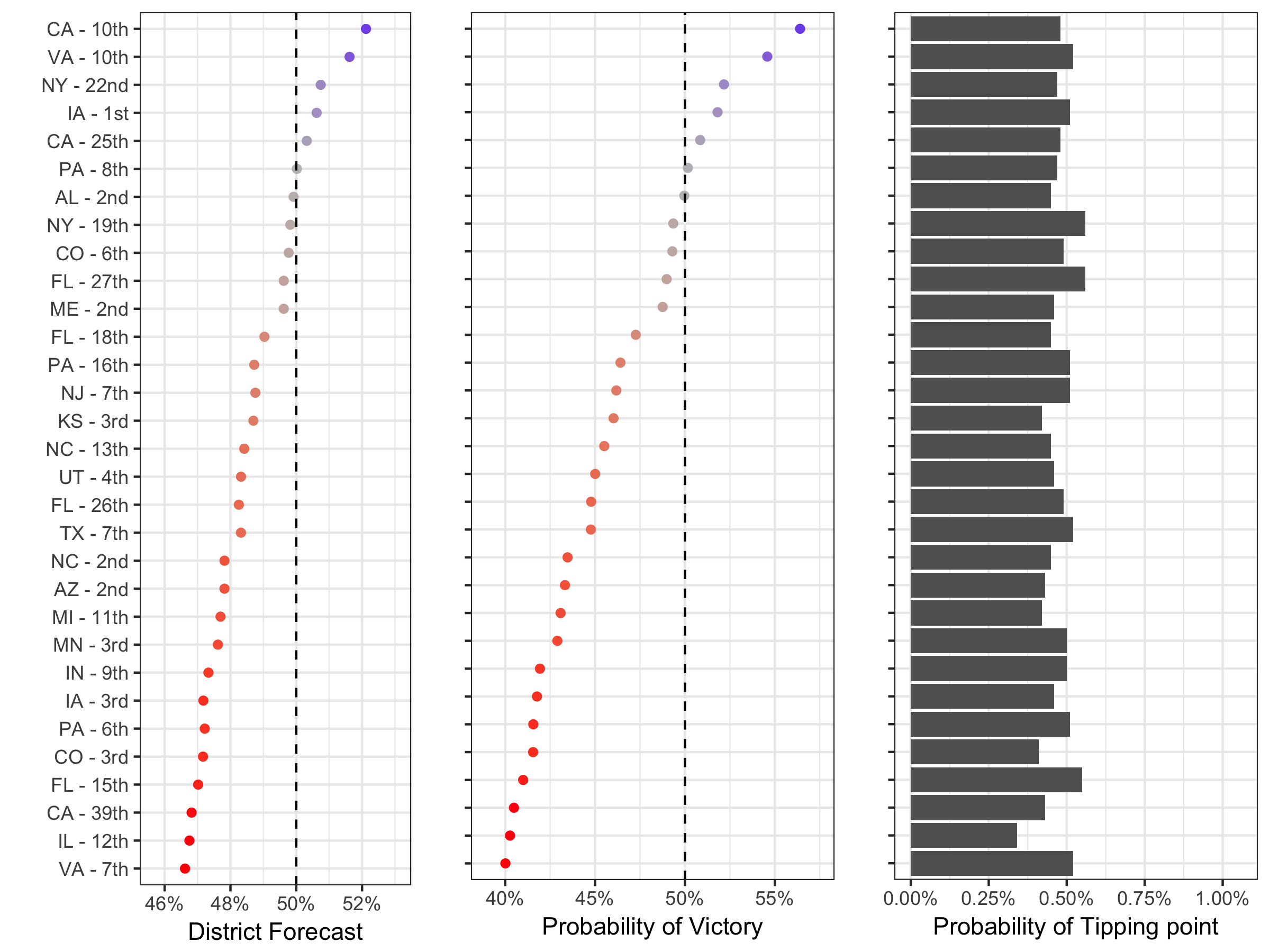
If the election were held today and assuming polling is not systematically flawed, there is a 100% chance that Democrats win over 50% of the popular vote, and more importantly, there is a 99% chance that they win a majority of seats in the House itself. However, we are still months away from the election. Simulations allow us to account for the variation we expect over time. Across 10,000 simulations of support **on election day**, Democrats win the popular vote in 62% of the simulations and a majority in the House in 55% of the simulations.

## Strategic targeting for 2018

Following [**Jackman (2014)**](https://www.cambridge.org/core/services/aop-cambridge-core/content/view/S1049096514000109), I apply national uniform swing to the 2016 two-way results of each Congressional district to identify key districts for targeting. About 30 districts have between a 40% and 60% chance of Democratic victory on election day, based on the national uniform swing. However, most of these key target seats are leaning towards Republicans; Democrats will need to win in over half of these to take back the House.



There are 198 seats where Democrats have more than a 60% chance of victory. 27 of these seats only lean towards the Democrat (less than an 80% chance of victory) and we must focus on protecting. An additional 91 seats lean Republican and are expansion seats if it looks like a wave year.



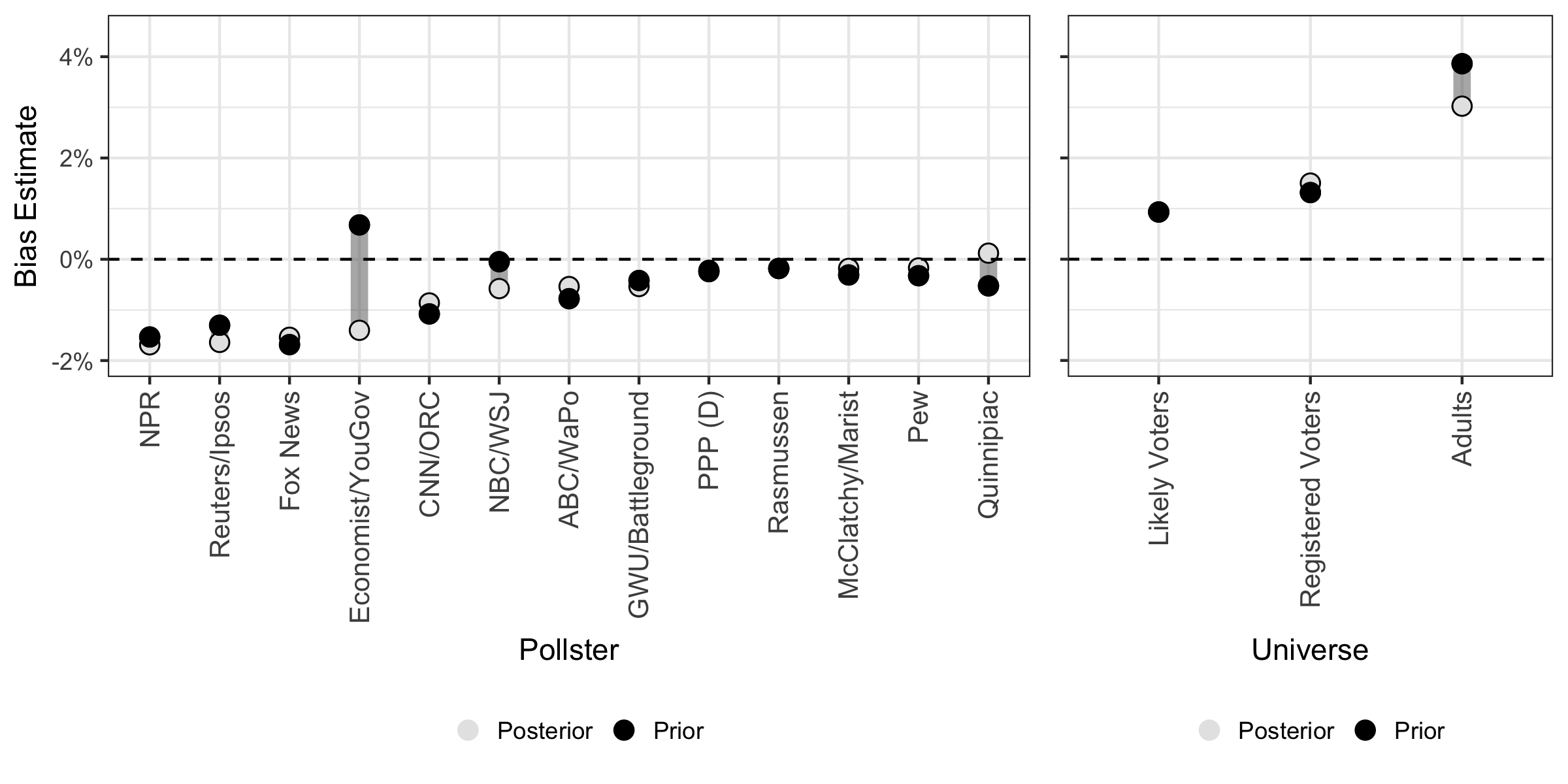
Importantly, since the election is still so far away, it’s not clear which of these districts is the most likely to be the 218th seat. In general, Democrats should allocate resources to so-called ‘tipping point’ districts. As seen in the third column above, this implies a fairly broad investment strategy since the election is so distant. Right now, 364 districts have some nonzero likelihood of being the tipping point on Election Day, but if the election were held today, just 20 districts would deserve investment. Until these key districts become clearer, **Democrats should fight everywhere**.

An alternative to uniform swing that would give more precise rankings now is building a forecast using the named-congressional ballot. While polling in sufficiently large numbers to produce accurate estimates would be costly, it would generate:

1. fine-grained, district-specific estimates of support that correctly rank the districts;
2. district-specific uncertainty allowing us to identify which districts are more likely to swing than others.

## Pollster and universe bias

Lastly, it is worth demonstrating that we observe only small changes in estimates of bias. Rasmussen was the reference pollster for this cycle, since its Bayesian prior had the least uncertainty (see Appendix A for more information) – so its estimate has been fixed. Only two pollsters changed the sign of their bias: The Economist/YouGov went from overestimating Democratic support to being more conservative vis-a-vis Rasmussen, and Quinnipiac moved in the opposite direction. On one hand, there is not much evidence that other pollsters are consistently more or less biased than our previous estimate; on the other, in some cases we have enough evidence to detect that two pollsters updated their methodologies enough to detect directional change in their biases.



Similarly, the estimate for bias of registered voter universes is similar across time: just about 1.5pp towards Democrats. On the contrary, holding the bias of likely voter universes constant, adult universes are considerably less biased than in previous cycles, but are still quite biased.

## The Data

The polling data were scraped from Real Clear Politics’ database for the [**2018**](https://www.realclearpolitics.com/epolls/other/2018_generic_congressional_vote-6185.html#polls) generic Congressional ballot. Each pollster has a slightly different wording (hence why we measure pollster bias), but they are all similar to: ‘If the elections for the U.S. House of Representatives were being held today, which party’s candidate would you vote for in your congressional district: The Democratic candidate or the Republican candidate?’ The named Congressional ballot question would account for incumbency effects and more closely mirror the choice voters will make in the voting booth. However, since not all candidates are known for 2018 yet, the generic ballot is the most common question being polled by public sources.

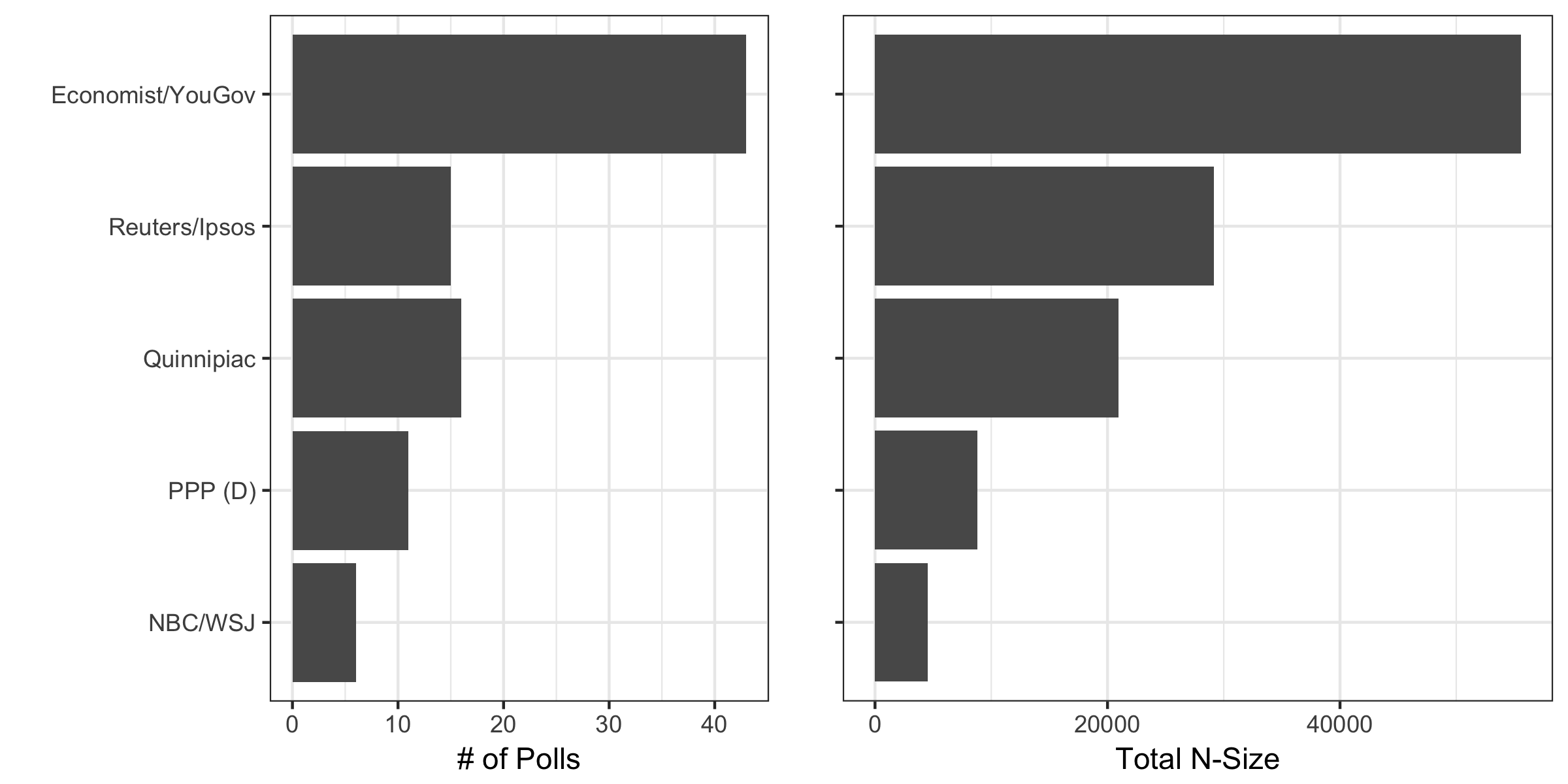
Only polls where the year, date range, pollster, sampling universe and sample size are all known were included.

Additionally, the polls’ results were transformed to reflect the two-way share for Democrats (Dem/(Dem+Rep)): it is a proportion between 0 and 1. Almost all Congressional races are only seriously contested by these two parties and one candidate only needs a plurality of support to win. Therefore, to make probability-of-victory estimates, we can just build one model that predicts the share of the two-party vote, reducing computation time significantly.

Time is transformed to be the rounded number of weeks between the middle day of the poll and election day. A daily model would be more precise, but would take more data to estimate with the same accuracy. Additionally, a weekly model acknowledges that most changes in support are transient and allows us to focus on meaningful changes. Lastly, the weekly model lowers the number of parameters that need to be estimated by a factor of 7, reducing computation time as well.

Only sampling universes and pollsters who have polled in at least one of the last six election cycles are included. This is because I use estimates of bias observed over the last elections to account for polling house effects. See more [**here**](https://github.com/nickahamed/BSTM/blob/master/proj/pollster_bias.html).

As of February 12, 2018, 105 polls from 13 pollsters contacting 133k respondents were used. These are the 5 largest pollsters. Full details and code available.



2016 U.S. House election results were taken from [**here**](https://github.com/Prooffreader/election_2016_data).

# Appendix A

To specify my random walk model, I follow [**Jackman (2005)**](http://eppsac.utdallas.edu/files/jackman/CAJP%2040-4%20Jackman.pdf). A given poll is assumed to be normally distributed with support as the mean and the standard deviation a function of and sample size. This would be specified as:

That poll is centered around mean , which itself is a function of , the true value of support at the time the poll was taken , , the bias of pollster , and , the bias of sampling universe . Fully specified, this is:

Due to the trends we see in our initial data exploration, a random walk model is appropriate. In such a model, support at time is normally distributed around support at time .

For these given specifications, I start with the following priors:

just follows the formula for variance of a sample. As a prior for the starting true value of support (), I use a uniform distribution over the minimum and maximum actual vote share of Democrats in the six elections analyzed. Lastly, as a prior for the true standard deviation (), I use a uniform distribution between 0 and 0.01. A value of 0.01 would reflect that 95% of week-to-week movement is within about 2pp in either direction, a fairly weak assumption. These priors are similar to [**Strauss (2007)**](http://www.mindlessphilosopher.net/princeton/strauss_reverse%20random%20walk.pdf).

Unlike [**Jackman (2005)**](http://eppsac.utdallas.edu/files/jackman/CAJP%2040-4%20Jackman.pdf) and [**Strauss (2007)**](http://www.mindlessphilosopher.net/princeton/strauss_reverse%20random%20walk.pdf), I have strong priors for and . Whereas they use a weak prior centered around 0, I use the best estimate from past election cycles and the observed uncertainty about that estimate. See more about the estimation of those [**here**](https://github.com/nickahamed/BSTM/blob/master/proj/pollster_bias.html), including the original priors and observed posteriors. However, using priors for all of the variables leaves the model underspecified. It is impossible to simultaneously estimate the bias for universe/pollster, and the true level of support each week. As such, I fix the prior for the pollster/universe for which we have the least uncertainty about at its bias estimate (set variance of prior = 0). Whereas past bias estimates were derived in relation to the true election result, this approach updates the priors in relation to this pollster/universe.